Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Using isotope composition and other node attributes to predict edges in fish trophic networks

Vyacheslav Lyubchich [*], Ryan J. Woodland

Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, USA

ARTICLE INFO

Article history:
Available online 6 June 2018

Keywords:
Deep neural network
Exponential random graph model
Food web
Node feature
Random forest
Random network

ABSTRACT

Stable isotope analysis becomes increasingly popular in ecological modeling. With exponential random graph models and machine learning techniques, this paper shows how predator isotope information and basic physical variables become predictors for the links in a trophic network.

1. Introduction

In recent years, increases in data availability and reductions in the costs of data collection and processing have expanded opportunities for the extensive analysis of ecological networks. At the same time, anthropogenic influence in natural systems makes it particularly important to study the robustness of ecological networks. Assessments of ecological networks can include examining the presence and predicting the occurrence of highly connected nodes and clusters, evaluating the response and resilience of large food webs to perturbations in their topology, and tracking the effect of biodiversity changes due to species introductions or removals (Aufderheide et al., 2013; Dunne et al., 2002; McCann, 2000; Namba, 2015; Vieira and Almeida-Neto, 2015).

Data for building such networks often come from observations of consumer foraging behavior (e.g., video footage) or by studying consumer stomach contents. In an aggregated form, these data become a food web of species (see Christensen and Walters, 2004; Kavanagh et al., 2004, and references therein on the methods for mass-balancing of aquatic trophic networks, e.g., Ecopath), however, individual observations allow us to study the networks in more detail. Individual or group-specific patterns in diet can result in consumers from the same species occupying different positions in a food web. The reasons include specifics of their habitat area, size-dependent foraging, prey abundance and diversity, and competitive interactions with other predators (including conspecifics).

Trophic networks of individuals are among the least studied forms of aquatic food web models. For fishes, stomach content data remain the most common source of information (Winemiller, 1990; Peterson et al., 2017). Sampling and identifying prey items from stomachs is labor intensive and requires specialized knowledge of prey taxonomy. Captured fishes often have empty stomachs, rendering them uninformative for such analysis. Moreover, stomach content information is inherently ‘noisy’ because it provides an instantaneous snapshot of the fishes’ ‘meal’ rather than a record of their diet preferences. To overcome the noisiness due to empty stomachs and possible mismatch of recent stomach contents with average diet, larger sample sizes are often required.

[*] Corresponding author.
E-mail addresses: lyubchich@umces.edu (V. Lyubchich), woodland@umces.edu (R.J. Woodland).

https://doi.org/10.1016/j.spl.2018.06.001

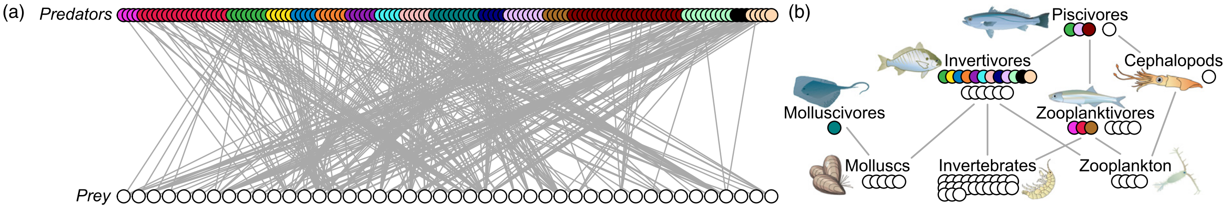

Fig. 1. Observed fish trophic network, based on stomach content: (a) bipartite graph of 132 individual fishes colored by species, 44 prey categories are white, (b) main trophic groups (the fauna images are courtesy of the Integration and Application Network, UMCES).

Alternatively, stable isotope composition of fish muscle tissues contains information about assimilated diet, reflecting time-integrated diet that is accessible regardless of current stomach contents or emptiness. Recent technological advances make stable isotope analysis fast and cost effective for studying fish networks (e.g., see Anderson and Cabana, 2007; East et al., 2017; Middelburg, 2014, and references therein). Whereas the referenced studies analyze networks of *species*, here, we demonstrate the application of stable isotope data and other easily measurable variables in analyzing trophic networks of *individual fishes*.

This article also demonstrates the way of implementing efficient machine learning techniques in random network analysis. Indeed, the potential of modern machine learning methods in predicting network connections has been largely untapped (Shi et al., 2015; Bianconi et al., 2009, and references therein). Here, we apply random forests and deep machine learning to forecast edges between nodes in a trophic network. We demonstrate our approach in comparison with exponential random graph models, using the real data on a marine fish community. We assess the results in three ways: by comparing the goodness-of-fit on a training dataset; by considering the network statistics not used in modeling, and by considering the individuals that were not used in modeling.

Section 2 describes the data, and Section 3—the methods. The main results are given in Section 4. A discussion in Section 5 concludes the paper.

2. Data

The data were collected from 1,271 fishes sampled by trawl in the Maryland inner continental shelf region of the Middle Atlantic Bight in July–September 2005, August 2007, and August 2008 as described by Woodland and Secor (2013). We use a subset of 164 fishes that had information on isotope composition: carbon (δ^{13}C) and nitrogen (δ^{15}N). All those individuals were identified to the species-level, measured for total length, weighed, and had their stomach contents classified using 44 different prey categories, including bony fishes (anchovies, flounders), crustaceans (shrimp, crab, and amphipods), zooplankton, mollusks, and polychaetes. About 22% of the individual fish weight records were missing and were calculated using the allometric regression: $Weight = A \cdot Length^{B} \xi$, where ξ is a multiplicative error term (Hayes et al., 1995). The estimates $\hat{A}$ and $\hat{B}$ for each species were obtained from Woodland and Secor (2013) and http://Fishbase.org.

We separate the fishes into two groups: those with non-empty stomachs (132 fishes of 17 species) and empty stomachs (32 fishes of 10 species). The ten species of the second group are a subset of the species from the first group. The first group is used to build a trophic network (Fig. 1a) and fit statistical models to this network (the design matrix has $n = 5808$ rows as the number of edges in a fully connected two-mode network). With those models, we can potentially infer the trophic connections (Fig. 1b) in the second group of fishes, where diet information from stomach contents was not available.

3. Methods

Let $G(V, E, A)$ be a random graph with sets of vertices V, edges E, their attributes A, and adjacency matrix $\mathbf{Y}$. In bipartite graphs, like the trophic network in Fig. 1a, V is represented by two disjoint sets V_1 and V_2. Elements of $\mathbf{Y}$, y_{ij}, are binary random variables taking on values 0 or 1 denoting, respectively, absence or presence of an edge between nodes $v_{1,i}$ and $v_{2,j}$ ($i = 1, \ldots, |V_1|; j = 1, \ldots, |V_2|$). We can match each element of $\mathbf{Y}$ with a p-vector of some observed network statistics $g(y)$ and use various methods for predicting the binary response (presence or absence of an edge) based on these covariates.

3.1. Exponential random graph models

We represent the distribution of $\mathbf{Y}$ using an exponential random graph model (ERGM, Handcock and Gile, 2010; Hunter et al., 2008):

$$\Pr_{\theta}(\mathbf{Y} = y) = h(y) \exp\left\{\theta^{\top} g(y) - \kappa(\theta)\right\}, \quad y \in \mathcal{Y}, \tag{1}$$

where $h(y)$ is the reference measure (for binary ERGMs, it is a Bernoulli-reference specifying probability of an edge present being 0.5), θ is the parameter vector ($\theta \in \mathbb{R}^p$), $\exp\{\kappa(\theta)\}$ is the normalizing constant for the exponential family of distributions, and $\mathcal{Y}$ is the support of **Y**. ERGMs use an approach analogous to generalized linear models for modeling the log-odds of observing a given network $G(V, E, A)$ given the set of network characteristics. The latter may include network statistics (e.g., number of edges, k-stars, or triangles; see Kolaczyk and Csárdi, 2014, and references therein for more details) as well as exogenous covariates defined by node attributes (such as isotope composition, species, or weight). Due to the nature of our research question (inferring the diet of individuals in a trophic network) and data used, the network information for an individual is not known upfront. Thus, we consider only the exogenous covariates, that also makes possible further comparisons with other methods using the same set of covariates.

3.2. Tree-based methods

From the variety of tree-based methods, we select three: the basic classification and regression trees (CART, Breiman et al., 1984), classic random forest (RF, Breiman, 2001), and case-specific random forest (CSRF, Xu et al., 2016). CART is a technique employing the predictors' values to recursively partition the set of corresponding responses into the groups of highest homogeneity (purity). In classification tasks, groups with all 0s or all 1s have the highest purity (often measured by the Gini index). With a large number of distinct values of predictors, it is possible to overfit data and separate an entire dataset into groups with only one observation per group (a case of high purity). To prevent overfitting, CART implements a complexity penalty based on the number of terminal nodes, however, a single tree can still be very unstable. Thus, rather than being a stand-alone procedure, more often CART becomes a basis for other methods, such as gradient boosting or random forests (Berk, 2016).

A RF consists of a large number of classification trees grown on bootstrap samples from the original data. In RF classifiers, the predicted class (existing or missing edge) for a data point X_0 is determined by the majority vote from all trees. Hastie et al. (2009) show that bootstrap aggregation (bagging) in RF does not change the bias, but reduces the variance of predictions compared to individual trees. To make trees less correlated with each other, only a random subset m out of p original variables is considered at each new tree split ($m \leqslant p$), with conventional choice of $m = \lfloor p/3 \rfloor$ (Hastie et al., 2009).

Xu et al. (2016) point out that in each random forest, some of the trees would be grown on bootstrap samples that are closer to X_0 than the rest. Indeed, if X_0 corresponds to some rare case, most of the trees (and their votes) would be irrelevant for predicting the class for X_0. Thus, Xu et al. (2016) suggest to upweight predictions from the trees that use data cases close to X_0. In their algorithm of CSRF, Xu et al. (2016) use bagging of trees in a random forest with $m = p$ and node size w to quantify the proximity as a ratio of the number of trees that put the training case i and X_0 together in one terminal node, to the total number of times it happens over all i ($i = 1, \dots, n$, where n is the size of the training dataset). The node size w becomes a tuning parameter that defines how concentrated the high weights are around the case X_0.

3.3. Deep neural networks

We construct a deep neural network (DNN, Hinton et al., 2006; Hinton and Salakhutdinov, 2006) for predicting the edges in a trophic network. DNNs are built by stacking restricted Boltzmann machines for supervised learning. Let $\mathbf{z}^{(l)}$, $\mathbf{b}^{(l)}$, and $\mathbf{y}^{(l)}$ be inputs, biases, and outputs of lth layer ($l = 1, \dots, L$), then the feed-forward operation using weights $\mathbf{w}^{(l)}$ for a standard DNN is:

$$z_i^{(l+1)} = \mathbf{w}_i^{(l+1)}\mathbf{y}^l + b_i^{(l+1)}, \quad y_i^{(l+1)} = f\left(z_i^{(l+1)}\right), \tag{2}$$

where $l = 1, \dots, (L-1)$, and $f(\cdot)$ is an activation function (Srivastava et al., 2014). In this study, we use a sigmoid function $f(x) = 1/(1 + \exp(-x))$ in a neural network of four layers total (one input and one output layer, two hidden layers) and backpropagation algorithm for fine-tuning the weights.

4. Main results

We start by fitting the network in Fig. 1a using variants of model (1) with different sets of covariates. Table 1 shows that adding more covariates improves the fit, however, the full model B3 has many coefficients that are not statistically significant. This may be due to an unbalanced species distribution and small group sizes: counts of fishes by species range from 3 (*Scophthalmus aquosus*) to 23 (*Pomatomus saltatrix*). Coefficients for all variables in models B1, B2, and B4 are individually statistically significant. Thus, we switch to model B4, which has only four parameters and a good fit (close to the full model, B3). Due to its parsimony, model B4 is also the preferred model based on BIC (Table 1).

In-sample performance of the tree-based and neural network classifiers is assessed in Table 2 using confusion matrices $\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$, where rows correspond to *true* missing and existing edges, columns—to *predicted* missing and existing network edges, giving the counts of true negatives (TN), true positives (TP), false negatives (FN), and false positives (FP). The overall misclassification rate is calculated for each method as: $R = n^{-1}(FP + FN)100\%$.

The actual proportion of existing edges is just 5.4% (316 out of 5808), thus, a naive assumption of the network being completely disconnected would lead to only a 5.4% error. To make the classifiers pick up the weak signal in this highly unbalanced dataset, we decrease CART's complexity parameter to 0.001 (Berk, 2016) and adjust case weights in RF to be four times higher when an edge is present (Wright and Ziegler, 2017). The CSRF in Table 2 uses $w = 10$ to define the

Table 1
Exponential random graphs models (ERGMs) fitted to the fish trophic network using approximate maximum likelihood, along with their number of parameters (p) and goodness-of-fit statistics, including Akaike and Bayesian information criteria (AIC and BIC).

Model	Covariates	p	Residual deviance	AIC	BIC
B1	*Length* + *Weight*	2	3639	3643	3656
B2	B1 + *Species*	18	2513	2549	2669
B3	B2 + δ^{13}C + δ^{15}N	20	2371	2411	2544
B4	B1 + δ^{13}C + δ^{15}N	4	2415	2423	2450

Table 2
Comparison of the classifiers on training data ($n = 5808$, $\|E\| = 316$).

Value	CART	RF	CSRF	DNN
Confusion matrix	$\begin{bmatrix} 5453 & 39 \\ 224 & 92 \end{bmatrix}$	$\begin{bmatrix} 5421 & 71 \\ 27 & 289 \end{bmatrix}$	$\begin{bmatrix} 5490 & 2 \\ 30 & 286 \end{bmatrix}$	$\begin{bmatrix} 5490 & 2 \\ 64 & 252 \end{bmatrix}$
R, %	4.5	1.7	0.6	1.1

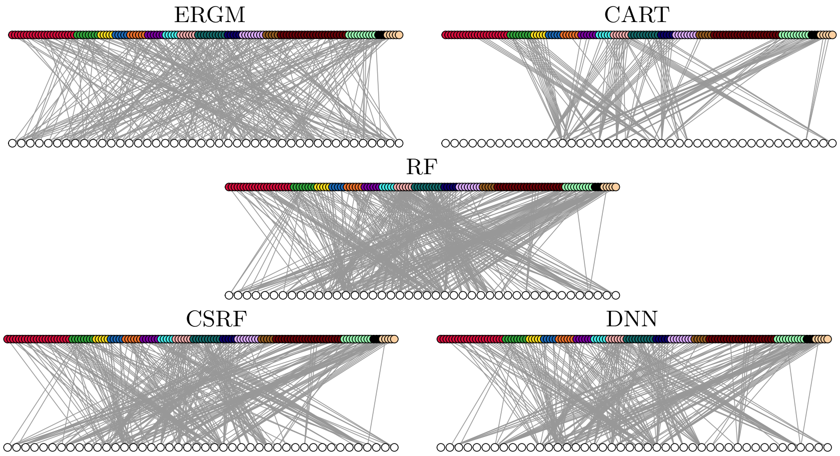

Fig. 2. Foraging networks based on predicted edges by different methods.

proximity of observations. The DNN comprises two hidden layers of 10 neurons each and uses 1,000 epochs for fine-tuning the weights using backpropagation, with the best fit achieved at the 710th iteration.

Among the trained classifiers, CSRF delivers the lowest overall misclassification rate, followed by DNN and classic RF. Due to the lack of an independent dataset for out-of-sample assessment of the methods' performance, we were forced to make comparisons based on the training data. Traditional cross-validation in this network context is problematic, because separating the dataset into training and testing would disrupt the whole observed trophic network. Fig. 2 shows fitted networks by all five methods, where the ERGM's network is one of the simulated networks from model B4. The confusion matrices in Table 2 and comparison of Fig. 2 with Fig. 1a show that the ERGM and CSRF fit the data well. The RF yields a network of higher density by over-predicting the total number of edges (360 vs. the observed 316). The DNN's trophic network has 254 predicted edges that are mainly concentrated at few prey categories, leaving many of them disconnected (cf. ERGM's and RF's networks in Fig. 2). The worst fit is delivered by CART, where the network shows very few prey categories being picked up (Fig. 2).

We also compare the methods using statistics that were not employed in the models, e.g., degrees of predator nodes, k. CSRF closely matches the observed degree distribution, followed by RF and, especially for $k > 2$, by DNN (Fig. 3). All methods overestimate the proportion of nodes without edges (the observed Pr(0) is exactly zero). The point estimates derived from CSRF and RF networks are close to the observed degree probabilities, while confidence intervals are readily available only from the ERGM simulations (for other ways of inference on network degree distribution, see Gel et al., 2017; Zhang et al., 2015).

In our last step, we use the trained models to predict prey categories for the second group of fishes, without stomach content data (empty stomachs), and use domain knowledge to assess the predictions. We used 13 such fishes that had their length, weight and isotope information within the species-specific range in the training dataset. The tree-based methods predicted at least one edge (non-zero node degree) for eight out of these 13 fishes—the edge connecting the new fishes with prey categories from the training dataset. All predicted edges agree with the current state of knowledge on shelf fish ecology.

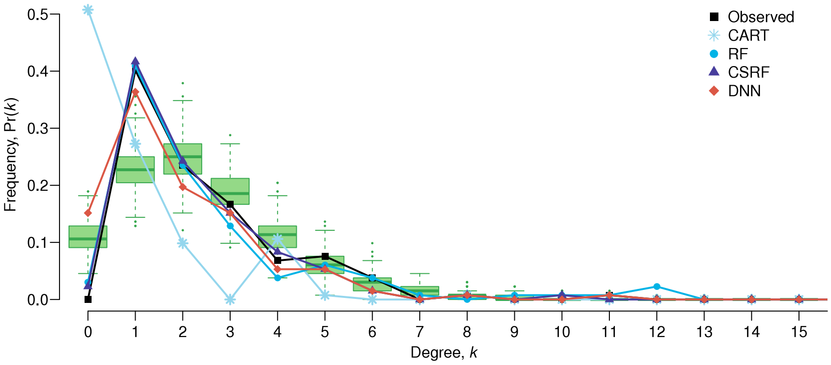

Fig. 3. Degree distributions from the observed and fitted trophic networks. Green boxplots correspond to 500 simulations from the ERGM B4.

5. Discussion

The considered statistical and machine learning models used easily measured covariates and showed varying success in reconstructing an empirical trophic network. One limitation of this study is that the models are evaluated on training data. The traditional approach of V-fold cross-validation is not a solution in this case, because removing some nodes would significantly alter the whole structure of the network. To evaluate the considered approaches more fairly, one would need a comparable dataset of an empirical trophic network for use as a testing sample.

We compared the performance of statistical and machine learning methods by assessing how well they fit the observed network. In our case study, the networks are constructed based on stomach contents information, which represent a snapshot of recent feeding by the fishes at the time of capture. Stomach contents analysis typically underestimates true trophic network density (mean degree) for many reasons. These include damage to prey resulting from ingestion and digestion that renders prey unidentifiable, natural patchiness of prey in the environment leading to spatiotemporal heterogeneity of stomach contents and different retention times of prey in fish stomachs (e.g., soft-bodied prey such as worms digest faster than hard-bodied prey such as crabs). The fishes in this dataset can be assumed to have more diverse diets than recorded, implying a denser network or even more substantial differences in degree distribution. Thus, the current overestimates of frequencies of higher degrees (e.g., $k > 11$ by random forest) can be considered natural for this dataset.

One caveat typical of stable isotope-based trophic studies is that prey, as well as the predators, must also be measured for stable isotope concentrations to parameterize mixing models (e.g., Parnell et al., 2010). Unfortunately, methods used to collect larger-bodied consumers such as fish for monitoring purposes are often inefficient at collecting prey. Our use of fish stable isotope values as covariates of trophic network models avoids the requirement of prey isotope values inherent in mixing models, suggesting that these data science methods could provide modeling opportunities for datasets in which comprehensive stable isotope datasets are not available for both predators and prey.

Acknowledgment

Lyubchich's research was supported in part by the US National Science Foundation under grant IIS 1633355.

References

Anderson, C., Cabana, G., 2007. Estimating the trophic position of aquatic consumers in river food webs using stable nitrogen isotopes. J. N. Am. Benthol. Soc. 26 (2), 273–285.
Aufderheide, H., Rudolf, L., Gross, T., Lafferty, K.D., 2013. How to predict community responses to perturbations in the face of imperfect knowledge and network complexity. Proc R Soc Lond [Biol] 280 (1773), 20132355.
Berk, R.A., 2016. Statistical learning from a regression perspective, second ed.. In: Springer Texts in Statistics, Springer, Switzerland.
Bianconi, G., Pin, P., Marsili, M., 2009. Assessing the relevance of node features for network structure. Proc. Natl. Acad. Sci. USA 106 (28), 11433–11438.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Breiman, L., Friedman, J.H., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press, New York.
Christensen, V., Walters, C.J., 2004. Ecopath with Ecosim: Methods, capabilities and limitations. Ecol. Model. 172 (2–4), 109–139.
Dunne, J.A., Williams, R.J., Martinez, N.D., 2002. Network structure and biodiversity loss in food webs: Robustness increases with connectance. Ecol. Lett. 5 (4), 558–567.
East, J.L., Wilcut, C., Pease, A.A., 2017. Aquatic food-web structure along a salinized dryland river. Freshwater Biol 62 (4), 681–694.
Gel, Y.R., Lyubchich, V., Ramirez Ramirez, L.L., 2017. Bootstrap quantification of estimation uncertainties in network degree distributions. Sci. Rep. 7, 5807.
Handcock, M.S., Gile, K.J., 2010. Modeling social networks from sampled data. Ann. Appl. Stat. 4 (1), 5–25.
Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed. Springer, New York.
Hayes, D.B., Brodziak, J.K.T., O'Gorman, J.B., 1995. Efficiency and bias of estimators and sampling designs for determining length-weight relationships of fish. Can J Fish Aquat Sci 52 (1), 84–92.

Hinton, G.E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18 (7), 1527–1554.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.

Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M., 2008. ergm: A package to fit, simulate and diagnose exponential-family models for networks. J. Stat. Softw. 24 (3), 1–29.

Kavanagh, P., Newlands, N., Christensen, V., Pauly, D., 2004. Automated parameter optimization for Ecopath ecosystem models. Ecol. Model. 172 (2–4), 141–149.

Kolaczyk, E.D., Csárdi, G., 2014. Statistical analysis of network data with R. In: Use R!, vol. 65, Springer, New York.

McCann, K.S., 2000. The diversity–stability debate. Nature 405 (6783), 228–233.

Middelburg, J., 2014. Stable isotopes dissect aquatic food webs from the top to the bottom. Biogeosciences 11 (8), 2357.

Namba, T., 2015. Multi-faceted approaches toward unravelling complex ecological networks. Popul. Ecol. 57 (1), 3–19.

Parnell, A.C., Inger, R., Bearhop, S., Jackson, A.L., 2010. Source partitioning using stable isotopes: Coping with too much variation. PloS One 5 (3), e9672.

Peterson, C.C., Keppeler, F.W., Saenz, D.E., Bower, L.M., Winemiller, K.O., 2017. Seasonal variation in fish trophic networks in two clear-water streams in the Central Llanos region, Venezuela. Neotrop. Ichthyol. 15 (2), e160125.

Shi, S., Li, Y., Wen, Y., Xie, W., 2015. Adding the sentiment attribute of nodes to improve link prediction in social network. In: P. 12th Internat. Conf. Fuzzy Systems and Knowledge Discovery, FSKD. pp. 1263–1269.

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Vieira, M.C., Almeida-Neto, M., 2015. A simple stochastic model for complex coextinctions in mutualistic networks: Robustness decreases with connectance. Ecol. Lett. 18 (2), 144–152.

Winemiller, K.O., 1990. Spatial and temporal variation in tropical fish trophic networks. Ecol Monograph 60 (3), 331–367.

Woodland, R.J., Secor, D.H., 2013. Benthic-pelagic coupling in a temperate inner continental shelf fish assemblage. Limnol Oceanogr 58 (3), 966–976.

Wright, M.N., Ziegler, A., 2017. ranger: A fast implementation of random forests for high dimensional data in C++and R. J. Stat. Softw. 77 (1), 1–17.

Xu, R., Nettleton, D., Nordman, D.J., 2016. Case-specific random forests. J. Comput. Graph. Statist. 25 (1), 49–65.

Zhang, Y., Kolaczyk, E.D., Spencer, B.D., 2015. Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. Ann. Appl. Stat. 9 (1), 166–199.